

Original Research

A Clinical Validity-Preserving Machine Learning Approach for Behavioral Assessment of Autism Spectrum Disorder

Abdulmalik A. Lawan ^{1,*}, Nadire Cavus ^{2,3}

1. Department of Computer Science, Kano University of Science and Technology, 713281 Wudil, Nigeria; E-Mail: aalawan@kustwudil.edu.ng
2. Department of Computer Information Systems, Near East University, 99138 Nicosia, Cyprus; E-Mail: nadire.cavus@neu.edu.tr
3. Computer Information Systems Research and Technology Centre, Near East University, 99138 Nicosia, Cyprus

* **Correspondence:** Abdulmalik A. Lawan; E-Mail: aalawan@kustwudil.edu.ng

Academic Editor: Raul Valverde

Special Issue: [Neuroscience and Information Technology](#)

OBM Neurobiology

2022, volume 6, issue 3

doi:10.21926/obm.neurobiol.2203138

Received: May 31, 2022

Accepted: September 20, 2022

Published: September 30, 2022

Abstract

Autism spectrum disorder (ASD) is a neuropsychiatric disorder associated with critical challenges related to social, communication, and behavioral issues. Recent studies have proposed machine learning (ML) techniques for rapid and accurate assessment of ASD. However, the mismatch between the ML techniques and the clinical basis of ASD assessment reduces the effectiveness of ML-based assessment tools. The present study proposed an approach that utilized the potential of ML modeling and preserved the clinical relevance of the assessment instrument used. Experimental results of the empirical scoring algorithm and multiple ML models employed revealed the possibility of achieving a clinically valid ML-based ASD assessment tool. This study provides a roadmap for real-life implementation of ML-based ASD screening and diagnostic tools that comprise few behavioral features and maintain clinical relevance.



© 2022 by the author. This is an open access article distributed under the conditions of the [Creative Commons by Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium or format, provided the original work is correctly cited.

Keywords

Autism spectrum disorder; screening; diagnosis; artificial intelligence; machine learning

1. Introduction

Autism spectrum disorder (ASD) is a lifelong neurodevelopmental disorder characterized by deficits in social interaction and communication and the presence of repetitive, restricted patterns of behavior, interests, or activities. In the United States, the recent prevalence of ASD is 1 in every 44 children [1], while globally, the prevalence of ASD is estimated to be 1.5% of the entire world population [2]. The global increase in the prevalence of autism necessitates the search for effective and early diagnostic processes for improved outcomes in socialization and communication and for guiding parents to adopt appropriate interventions for their children with ASD [3-7]. Several studies have proposed the implementation of machine learning (ML)-enabled systems for rapid and cost-effective assessment of the disorder [8-11].

The promising outcomes achieved through the application of ML techniques across several research endeavors have motivated the increasing application of ML in assessing ASD on the basis of either genetic, physical biomarker, brain imaging, or behavioral data [12-14]. Another stem of behavioral studies supported by computer-assisted technologies relies on the use of physical biomarkers for assessing stereotypical and repetitive behaviors in people with ASD based on movement sensors and other vision-based technologies for facial movement, body movement, and eye gaze tracking, among others [15-18]. Specifically, the present research is aligned with studies that generate behavioral data based on questionnaire instruments and use ML algorithms to model the behavioral data. However, despite the excellent evaluation metrics achieved in ML-based behavioral studies, it is evident that the research methods and the generated ML models could lead to inaccurate assessments by professionals. For instance, in addition to improving diagnostic accuracy, most of the studies focused on reducing or transforming the items of the assessment instruments by using various data-centric approaches [12]. However, several studies did not investigate the relevance of the data-centric approaches, the sufficiency of the modeling parameters, and the resulting ML models themselves against the basic assumptions of clinical assessment of ASD symptoms [12, 19].

The present study, therefore, aimed to investigate the advances in the application of ML techniques to the behavioral assessment of patients with ASD and proposed a novel ML-based approach that preserves the clinical validity of the screening and diagnostic tool by adhering to the conceptual foundation followed by professionals to assess ASD. A comparative analysis was conducted on the classification performance of the empirical scoring algorithm and various ML models based on different experimental scenarios that align/misalign with the clinical approach. Accordingly, the proposed approach utilized the advantages of ML techniques while preserving the clinical validity of the assessment instrument by adhering to the clinical procedures used by professionals in administering the diagnostic tool.

2. Related Research

ML techniques are being increasingly applied for rapid and cost-effective assessment of ASD based on various datasets that contain data related to genetics, brain imaging, or behavioral symptoms [12, 14]. Behavioral symptoms are collected through various modalities, including questionnaire-based items and physical biomarkers. Unlike questionnaire-based symptoms, physical biomarkers such as facial movement, body movement, and eye gaze are evaluated by sensor-based and vision-based technologies [15-18]. For instance, Kowallik and Schweinberger [17] reviewed 36 sensor-based studies for improving ASD assessment and intervention. The reviewed studies focused on sensory inputs from various biomarkers, including voice, face, and body movements. Recently, Negin et al. [18] proposed a vision-based approach by using a novel video dataset of human actions to recognize stereotypic ASD behaviors. The predictive study achieved the best results by using various ML classifiers based on multilayer perceptron, Naive Bayes, and support vector machines together with other data-centric frameworks. Both studies revealed the high potential of sensor-based and vision-based behavioral assessment of ASD.

In particular, the present study is aligned with several studies that utilized questionnaire-based behavioral data for ML modeling. Descriptive analyses of the research area indicated that streamlining the data collection instruments using various dimensionality reduction approaches of feature selection and transformation is the common data pre-processing activity reported in previous studies [12, 14]. Specifically, many studies aimed at streamlining the assessment instruments by reducing the dimension of the datasets, followed by ML data modeling and performance evaluation on the reduced datasets [12, 20-24]. Some of the common dimensionality reduction techniques used are Trial-Error Feature Selection [10, 21, 24-26], Variable Analysis (Va) [27, 28], Chi-Square testing (CHI) and Information Gain (IG) [29], Correlation-Based Feature Selection (CFS) [11], and ML-Based Feature Selection [23, 30]. Furthermore, data-centric studies conducted ML modeling using various ML algorithms such as Random Forest [11, 27, 31, 32], Support Vector Machines (SVMs) [22, 23, 30, 33, 34], Decision Trees [21, 24-26], and Logistic Regression [10, 22, 29].

The data collection or assessment instruments form the basis of behavioral studies on ASD symptoms. Previous studies used retrospective datasets of various assessment instruments such as Autism Quotient (AQ) [9, 10, 27-29, 31, 32, 34-36], Q-CHAT [9, 28, 35, 36], Autism Diagnostic Observation Schedule (ADOS) [21-23, 26, 30, 37], Autism Diagnostic Interview-Revised (ADI-R) [24, 33, 37], and Social Responsiveness Scale (SRS) [33, 38, 39]. Accordingly, the most utilized sources of the datasets include Autism Genetic Resource Exchange, Boston Autism Consortium, Simons Simplex Collection [21, 22, 24, 26, 30, 38, 39], National Database for Autism Research (NDAR) [21, 22], Simons Variation In Individuals Project (SVIP) [21, 22, 30], and UCI ML repository [9, 10, 27-29, 31, 32, 34-36].

Apart from the common aim of dimensionality reduction and ML modeling, other previous studies focused on ML algorithm optimization [32, 34], input optimization [27, 28, 31, 35], and implementation of ML-based screening applications (apps) [9, 11]. For instance, Goel et al. [32] improved the performance of a Random Forest classifier by using a proposed Grasshopper Optimization Algorithm. The modified classifier outperformed commonly used ML models by predicting ASD with almost 100% accuracy, specificity, and sensitivity. Similarly, Suresh Kumar and Renugadevi [34] investigated input optimization by using the Differential Evaluation (DE) algorithm.

The proposed DE optimized SVM parameters and achieved superior performance over commonly used SVM and artificial neural network (ANN); the DE also optimized ANN for the correct classification of ASD cases. Other studies focused on the comparative performance of various dimensionality reduction techniques. For instance, Thabtah et al. [28] reported a comparative evaluation of Va, IG, and CHI in decreasing AQ items where Va outperformed other techniques by deriving fewer items that lead to excellent ML models. Pratama et al. [27] replicated this study and recorded higher sensitivity and specificity values of 87.89% in AQ-Adults with RF and 86.33% in AQ-Adolescents with SVM, respectively. Despite the superior performance metrics reinforced by the dimensionality reduction techniques, none of the preceding studies justified the conformity of the data-centric techniques with the conceptual foundation for the clinical diagnosis of ASD. Furthermore, because of the absence of standardized medical tests for numerical quantification of ASD [2], clinical assessment of the disorder relied on the careful application of the common diagnostic scales based on human knowledge and experience. Accordingly, ML-based studies must balance the trade-off of streamlining behavioral scales on the one hand and implementing clinically valid diagnostic systems on the other hand. In other words, implementing valid scales that adequately cover the human knowledge for the clinical diagnosis of ASD is critical to the real-life deployment of ML-based tools [12]. Thus, innovative approaches that could be tracked by professionals based on clinical relevance are required.

Several challenges impede the real-life deployment of ML-based ASD screening and diagnostic tools [12, 14, 40]. Specifically, the good performance of the data-centric approaches based on the commonly used evaluation criteria cannot ensure the clinical relevance of the resulting ML models. Accordingly, the commonly used performance metrics of specificity, sensitivity, and classification accuracy cannot adequately capture the human knowledge employed by professionals in identifying behavioral symptoms of ASD. Therefore, promising studies on the real-life deployment of ML-based ASD assessment systems must be supported by a clear understanding of the clinical foundation of the screening and diagnostic tools and the logical concepts of the data-centric techniques. Specifically, none of the previous studies aimed to preserve the clinical validity of the assessment instruments. The novelty of the present study is the preservation of the clinical validity of the assessment instrument while benefitting from the precision of the ML algorithms employed. The present study retained all the items of the data collection instrument and treated each item as an integral part of computing a few clinically valid input parameters. It also explored a novel data intelligence technique that accomplished both excellent performance metrics and conformity with the conceptual basis for the clinical diagnosis of ASD.

3. Methodology

3.1 Proposed Research Methodology

The primary aim of the present study is to demonstrate the predictive performance of various ML models on a novel screening instrument and propose a promising approach that ensures rapid and accurate screening of patients with ASD while preserving the clinical validity of the screening instrument. For this purpose, a scientific procedure was carefully implemented to achieve the research aim and objectives, as shown in Figure 1.

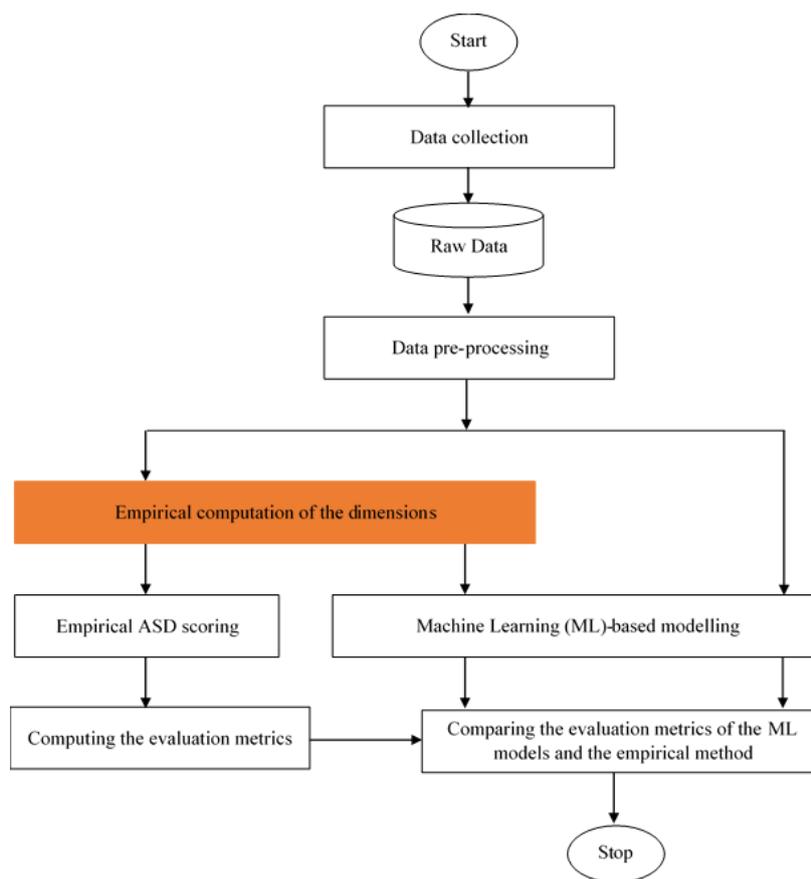


Figure 1 Flowchart of the proposed research methodology.

3.2 Participants

The study data were collected through web-based and printed questionnaires administered to voluntary caregivers, parents, and other relatives of children who were diagnosed to have neurodevelopmental disorders, including ASD, based on a purposive sampling approach. Some of the control cases were, however, drawn from participants with no symptoms of ASD or no comorbid neurodevelopmental disorders. Nonetheless, because of the lack of direct access to a sufficient number of patients with ASD through parents and caregivers, some of the responses were collected from teachers and clinicians of children with ASD. By using both data collection approaches, 411 anonymized responses were obtained. Cases with missing values were eliminated, which reduced the number of responses to 380 valid cases containing 171 ASD cases and 209 controls.

3.3 Data Collection Instrument

The data collection instrument named Child Development for Household Survey to Estimate Burden of ASD (CDHSEBA) is a questionnaire with an empirical scoring algorithm for assessing children “at-risk” of ASD. CDHSEBA can be used by parents, caregivers, clinicians, and researchers to screen ASD symptoms in children. Researchers at the Childhood Neuropsychiatric Disorders Initiative (<https://cndinitiatives.org/>) developed this instrument with its empirical scoring algorithm on the basis of the diagnostic criteria described in DSM-5. The empirical scoring algorithm provides logical and numerical measures for ASD symptoms. Sensitivity, specificity, and classification accuracy are the commonly used evaluation metrics for confirming the scientific rigor of a diagnostic

instrument in health-related studies [41]. In the present study, the data collection instrument achieved a high sensitivity of 97%, with classification accuracy and specificity of 56% and 23%, respectively. Figure 2 shows the procedure for computing the ratings of ASD symptoms. The rating scale is based on 0 and 1; if the response is NO (i.e., the behavior being probed is not present), it is coded as 0, while if the response is YES (i.e., the behavior being probed is present), it is coded as 1. The total score for the symptoms is then calculated, and YES or NO decision is provided on each section of the questionnaire. Consequently, the overall decision is computed by following the criteria for the diagnosis of ASD given in DSM-5, as shown in Figure 2; this summarizes how the empirical scoring algorithm validates whether the questionnaire responses meet the two conditions for “at-risk” ASD.

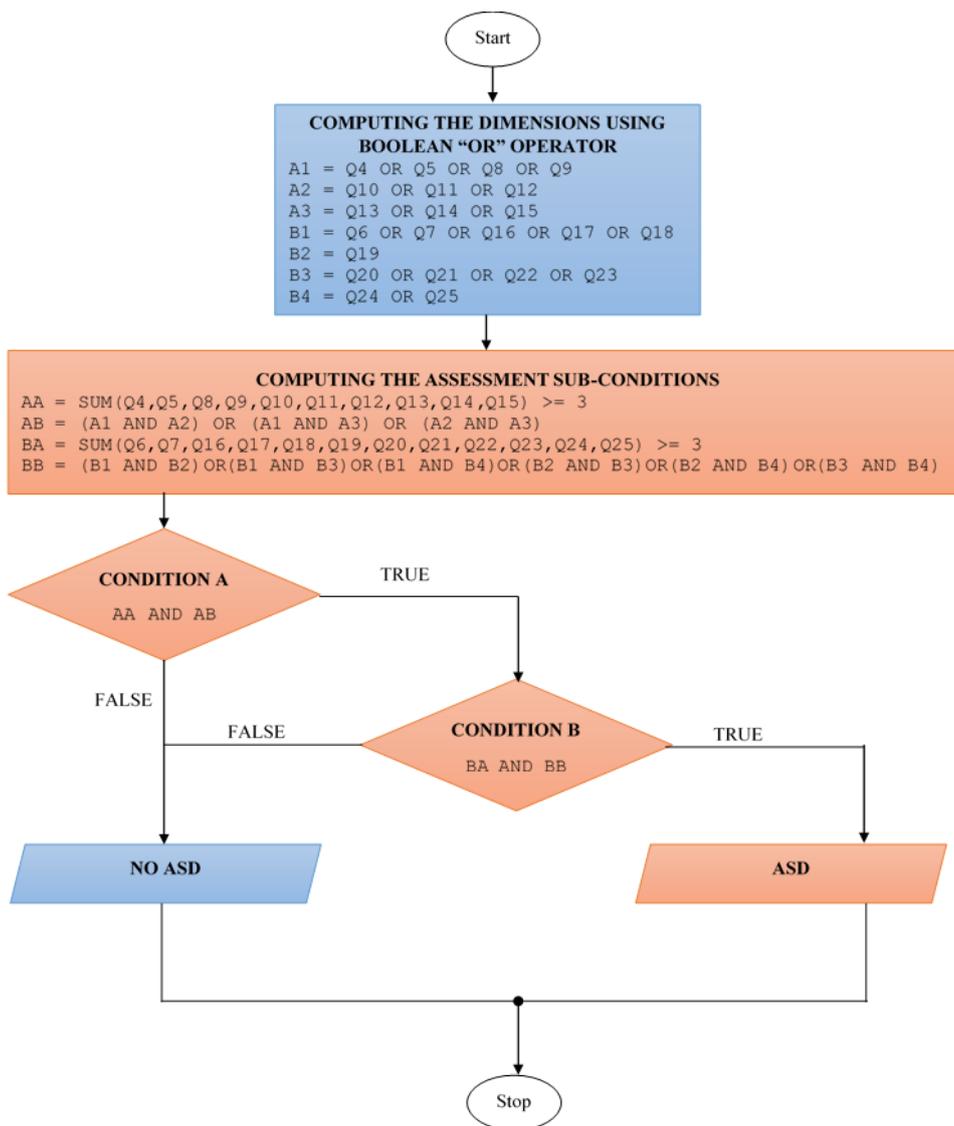


Figure 2 Flowchart of the empirical scoring algorithm of the study.

The proposed questionnaire contains less than 30 items (Appendix A) upon which the symptoms of ASD are scored, and the scoring algorithm follows section-by-section computations to meet the DSM-5 diagnostic criteria. Part 1 of the questionnaire captures demographic information (i.e., items 1, 2, and 3), while part 2 is categorized into sections A and B, described as follows.

3.3.1 Deficits in Social Communication

This section of the questionnaire contains items 4, 5, 8, 9, 10, 11, 12, 13, 14, and 15, which cover deficits in social communication and can be further grouped into three major categories in accordance with the DSM-5 criteria:

A1: Deficits in socio-emotional reciprocity (items 4, 5, 8, and 9)

A2: Deficits in nonverbal communication (items 10, 11, and 12)

A3: Deficits in developing, maintaining, and understanding relationships (items 13, 14, and 15)

Condition A: The patient is said to be presenting with social communication deficits if they receive a score of YES in 3/10 of these symptoms, and the symptoms must be from at least two different categories, i.e., the response must have a YES in at least A1 and A2, A1 and A3, or A2 and A3.

3.3.2 Restricted Behavior

This section of the questionnaire captures information on the presence of restricted and repetitive patterns of behavior, activities, or interests. Items in the questionnaire covering these aspects are 6, 7, 16, 17, 18, 19, 20, 21, 22, 23, 24, and 25, and these items can be further grouped into four major subcategories in accordance with the DSM-5 criteria:

B1: Stereotyped movements, language, or use of speech (items 6, 7, 16, 17, and 18)

B2: Insistence on sameness and inflexibility thought (item 19)

B3: Highly restricted, fixated interests and abnormal intensity in focus (items 20, 21, 22, and 23)

B4: Sensitivity to sensory input (items 24 and 25)

Condition B: The patient is said to be presenting with repetitive and stereotyped behavior if they present with 3 of the above-listed symptoms, with the symptoms being elicited from two different subcategories, i.e., a combination of the positive screen in B1 and B2, B1 and B3, B1 and B4, B2 and B3, B2 and B4, or B3 and B4.

In the present study, CDHSEBA was chosen because of the following three reasons that align it with the aim and settings of the research. First, CDHSEBA has fewer items, which meets the requirement for a rapid screening instrument with fewer items than the common gold standards. Second, the clinical empirical scoring method of the data collection instrument involves some form of dimensionality reduction; specifically, it has customized rules for feature transformation, in which the complete set of items are transformed into fewer dimensions (i.e., A1, A2, A3, B1, B2, B3, and B4) that subsequently lead to the main conditions upon which at-risk ASD cases are identified. Third, the data collection instrument is being used in an environment similar to the data collection units. Thus, there will be little or no environmental effect on the interpretability of the study findings.

3.4 Data Analysis

For data analysis, SPSS 25, Microsoft Excel 2016, and MATLAB R2019b were used. Before the ML modeling of the collected data, additional variables were computed using the data transformation feature of SPSS. The SPSS syntax used for computation is shown in Table 1.

Table 1 SPSS syntax for computing the variables.

```
GET FILE = 'C:\Users\hp\Google Drive\Researches\PHD CIS\Thesis report\Data\Data.sav'.
DATASET NAME DataSet1 WINDOW = FRONT.
COMPUTE Q18 = Q18A | Q18B | Q18C | Q18D | Q18E.
EXECUTE.
COMPUTE Q24 = Q24A | Q24B.
EXECUTE.
COMPUTE A1 = Q4 | Q5 | Q8 | Q9.
EXECUTE.
COMPUTE A2 = Q10 | Q11 | Q12.
EXECUTE.
COMPUTE A3 = Q13 | Q14 | Q15.
EXECUTE.
COMPUTE B1 = Q6 | Q7 | Q16 | Q17 | Q18.
EXECUTE.
COMPUTE B2 = Q19.
EXECUTE.
COMPUTE B3 = Q20 | Q21 | Q22 | Q23.
EXECUTE.
COMPUTE B4 = Q24 | Q25.
EXECUTE.
COMPUTE condition AA = (Q4 + Q5 + Q8 + Q9 + Q10 + Q11 + Q12 + Q13 + Q14 + Q15) >= 3.
EXECUTE.
COMPUTE condition AB = (A1 & A2) | (A1 & A3) | (A2 & A3).
EXECUTE.
COMPUTE condition BA = SUM (Q6, Q7, Q16, Q17, Q18, Q19, Q20, Q21, Q22, Q23, Q24, Q25) >= 3.
EXECUTE.
COMPUTE condition BB = (B1 & B2) | (B1 & B3) | (B1 & B4) | (B2 & B3) | (B2 & B4) | (B3 & B4).
EXECUTE.
COMPUTE condition A = condition AA & condition AB.
EXECUTE.
COMPUTE condition B = condition BA & condition BB.
EXECUTE.
COMPUTE computed ASD status = condition A & condition B.
EXECUTE.
COMPUTE TP = (clinical Status = 1) & (computed ASD status = 1).
EXECUTE.
COMPUTE TN = (clinical Status = 0) & (computed ASD status = 0).
EXECUTE.
COMPUTE FP = (clinical Status = 0) & (computed ASD status = 1).
EXECUTE.
COMPUTE FN = (clinical Status = 1) & (computed ASD status = 0).
EXECUTE.
```

Here, Q18 is the summarized value derived from the sub-items Q18A, Q18B, Q18C, Q18D, and Q18E using the OR (i.e., |) Boolean operator. Similarly, Q24 was computed based on Q24A and Q24B. As shown in the code listing, the seven sub-dimensions of the data collection instrument (i.e., A1, A2, A3, B1, B2, B3, and B4) were equally derived on the basis of their corresponding items. Nonetheless, the sub-conditions for assessing the disorder were correspondingly computed as condition AA, condition AB, condition BA, and condition BB. Here, condition AA tests whether at least three responses were YES on the items under section A of the data collection instrument, while condition AB confirms whether the YES responses were from either of the combinations of items under A1 and A2, A1 and A3, or A2 and A3, as explained while describing the manual scoring algorithm. Similarly, condition BA tests whether at least three responses from the items in section B of the data collection instrument were YES, while condition BB confirms whether the YES items were from either of the combinations B1 and B2, B1 and B3, B1 and B4, B2 and B3, B2 and B4, or B3 and B4, as described in the manual scoring algorithm.

The code listing showed the computation of the main conditions for diagnosing the disorder (i.e., condition A and condition B), where condition A was computed as TRUE if both condition AA and condition AB were TRUE, and condition B was computed similarly based on condition BA and condition BB. Finally, the code listing captured the key variable used in identifying the screening status of the participants (i.e., computed ASD status). Accordingly, the computed ASD status was TRUE if both conditions A and B were TRUE. Additionally, the basic evaluation metrics are shown in the code listing. Specifically, true positive (TP), true negative (TN), false positive (FP), and false negative (FN) were identified according to the computed status (i.e., computed ASD status) and the previous status (i.e., clinical status) as indicated in the questionnaire response. Evaluation metrics of the manual scoring algorithm were computed on the basis of the computed TP, TN, FP, and FN values. TP is the number of patients already diagnosed with ASD, and the screening instrument also classified them as ASD positive. FP is the number of patients that are truly non-autistic (i.e., belonging to the control group), but the screening instrument classified them as ASD positive. FP is also called a Type-I error. TN is the number of patients that are truly non-autistic (i.e., belonging to the control group), and the screening instrument also classified them as ASD negative. FN is the number of patients already diagnosed with ASD, but the screening instrument classified them as ASD negative. FN is also called a Type-II error.

3.4.1 Sensitivity and Specificity Analysis of the Manual Scoring Algorithm

Sensitivity and specificity are statistical measures that indicate the predictive value of an instrument in classifying positive and negative cases in a test [42, 43]. In the present study, the data on the predictive performance of the manual scoring algorithm are presented using a confusion matrix, while assorted formulae were followed to provide the commonly used evaluation metrics. The predictive performance of the scoring algorithm depicted in Figure 2 is, however, empirical and based on linear equations. ML models are usually developed to validate the empirical findings derived from manual scoring algorithms. Notably, previous studies have shown the improved accuracy of ML algorithms over manual scoring algorithms [29, 31, 33]. However, some of the previous studies that used various data pre-processing techniques did not preserve the clinical validity of the data collection instrument [12, 19]. Accordingly, in the present study, an alternative approach to manual scoring was used based on the ML algorithms. Specifically, to provide

comparative findings, the present study used both linear and nonlinear ML classification algorithms to capture possible nonlinear patterns in the data and to evaluate the performance of the models in classifying “at-risk” ASD cases without compromising the conceptual validity of the data collection instrument. The technique proposed in the present study grouped items of the collection instrument into distinctive dimensions that align with the use of human knowledge in the clinical assessment of ASD. Thus, the derived dimensions were used in training the ML models. Various data scenarios with a reduced and extended list of items were experimented with to provide more comparative results.

3.4.2 Sensitivity and Specificity Analysis of the ML Models

Model development in ML is a data-centric process that involves training the model with one part of the data and testing with the other part. Figure 3 shows the workflow diagram for constructing and evaluating the multiple ML models.

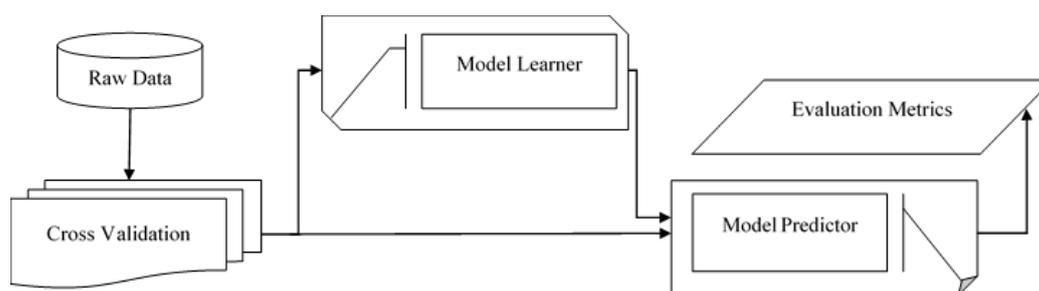


Figure 3 Machine learning-based classification of “at-risk” autism cases for the study.

The modeling approach began by defining and presenting both input and target parameters from the raw data, followed by iterative data resampling using 10-fold cross-validation. Accordingly, in each of the 10 iterations employed, the model learner and the model predictor were used in training and testing the ML models. Subsequently, evaluation metrics from the 10 cross-validations and modeling stages were averaged for comparative evaluation of the performances of the models.

3.5 Experimental Setting

This section presents the experimental settings for the comparative analysis of the ML algorithms and the empirical scoring algorithm. While the empirical scoring algorithm used the items described under the data collection tool (i.e., Q4–Q25), multiple ML algorithms were implemented using a different combination of the CDHSEBA raw and processed parameters. Specifically, 25 ML algorithms were implemented according to four different data scenarios. Data Scenarios 1 and 3 involved the raw items of the CDHSEBA, while Scenarios 2 and 4 contained the transformed CDHSEBA dimensions, as described in Table 2. These scenarios were meant to provide comparative results on the impact of the clinical data transformation on the performance of the ML algorithms based on the commonly used evaluation metrics and weigh the results against the trade-off of preserving the clinical validity of the data collection tool as well as the developed ML models. The study analyzed multiple ML algorithms because each algorithm has a different learning style in processing the dataset [42].

Table 2 Highlights of the different data scenarios for the development of models.

Data scenario	Input variables	Target variable	Description
Scenario 1	Q4, Q5, Q6, Q7, Q8, Q9, Q10, Q11, Q12, Q13, Q14, Q15, Q16, Q17, Q18, Q19, Q20, Q21, Q22, Q23, Q24, Q25	clinical Status	The inputs utilized were similarly used in the manual scoring method.
Scenario 2	A1, A2, A3, B1, B2, B3 and B4	clinical Status	The inputs used are the dimensions derived by categorizing the items in scenario 1 and averaging the responses.
Scenario 3	Q4, Q5, Q6, Q7, Q8, Q9, Q10, Q11, Q12, Q13, Q14, Q15, Q16, Q17, Q18A, Q18B, Q18C, Q18D, Q18E, Q19, Q20, Q21, Q22, Q23, Q24A, Q24B, Q25, and Q26 (severity level)	clinical Status	In this scenario, the complete set of items was passed to the machine learning algorithms without any processing or categorization. Unlike scenario 1, Q26 (severity level) was included, and the sub-items in Q18 and Q24 items were not averaged.
Scenario 4	A1, A2, A3, B1, B2, B3, B4, and Q26 (i.e., severity level)	clinical Status	This is similar to scenario 2 with the addition of demographic variable of severity level (i.e., Q26)

The proposed empirical scoring algorithm was implemented on SPSS version 25, as shown in Table 1. The variable computation function of SPSS was used for implementing the empirical scoring algorithm. The multiple ML classifiers were implemented on MATLAB version R2019b. The classification learning module of the MATLAB package was used to train the ML models. In testing the models generated by the 25 ML algorithms, 10-fold cross-validation was adopted in each of the four experimental data scenarios. Therefore, in each of the 10 cross-validations, the training dataset was partitioned into 10 subsets. The remaining nine data subsets were randomly used by the classification algorithm for testing the classifier. This validation process was iterated 10 times before averaging the classification error rates. Moreover, no hard coding was performed as the algorithm module, and the cross-validation procedures were embedded in the MATLAB R2019b platform and were selected from the graphical user interface before the learning phase. Finally, all the experimental runs were conducted on a personal computer with Microsoft Operating System. The different parameter combinations used in each experimental scenario are described in Table 2.

4. Results and Discussion

4.1 Confusion Matrix of the Empirical Scoring Algorithm

The basic parameters of TP, FP, TN, and FN, highlighted with the help of Table 3, were used for deriving different evaluation metrics, including classification accuracy, specificity, and sensitivity, to evaluate the performance of both empirical and ML algorithms.

Table 3 Confusion matrix of the empirical scoring algorithm.

	Predicted class POSITIVE ("at-risk" of ASD)	Predicted class NEGATIVE (control group)
Actual class POSITIVE ("at-risk" of ASD)	166	5
Actual class NEGATIVE (control group)	162	47

(TP: true positive, FN: false negative, FP: false positive, TN: true negative)

From the data (i.e., TP = 166, FN = 5, FP = 162, TN = 47), by using the formulae Sensitivity = TP/(TP + FN), Specificity = TN/(TN + FP), and accuracy = (TP + TN)/(TP + TN + FP + FN), the percentage sensitivity, specificity, and accuracy were estimated to be 97.1%, 22.5%, and 56.1%, respectively.

4.2 Comparative Performance of the ML Models and the Empirical Scoring Algorithm Across Various Experimental Scenarios

Tables 4–7 summarize the TP, FP, TN, and FN rates achieved by the multiple ML models implemented under data Scenarios 1, 2, 3, and 4, respectively. Comparative analysis showed that the empirical scoring algorithm outperformed the multiple ML models across all the experimental scenarios by achieving the highest sensitivity of 97%. However, in the first experimental scenario, Fine Gaussian SVM exhibited the highest specificity of 99% with the lowest classification accuracy of 57% (only 1% ahead of the empirical scoring algorithm) (Table 4). Overall, Coarse Gaussian SVM and Ensemble Bagged Trees achieved the highest accuracy of 78% in this scenario.

Table 4 Modeling results for Scenario 1.

ID	Model Name	TN	FP	FN	TP	Sensitivity	Specificity	Accuracy
1	Fine Tree	142	67	59	112	0.65	0.68	0.67
2	Medium Tree	144	65	53	118	0.69	0.69	0.69
3	Coarse Tree	139	70	71	100	0.58	0.67	0.63
4	Linear Discriminant	161	48	42	129	0.75	0.77	0.76
5	Quadratic Discriminant	0	0	0	0	0	0	0
6	Logistic Regression	163	46	46	125	0.73	0.78	0.76

7	Gaussian Naïve Bayes	0	0	0	0	0	0	0
8	Kernel Naïve Bayes	177	32	57	114	0.67	0.85	0.77
9	Linear SVM	171	38	50	121	0.71	0.82	0.77
10	Quadratic SVM	165	44	51	120	0.7	0.79	0.75
11	Cubic SVM	156	53	58	113	0.66	0.75	0.71
12	Fine Gaussian SVM	207	2	160	11	0.06	0.99	0.57
13	Medium Gaussian SVM	174	35	52	119	0.7	0.83	0.77
14	Coarse Gaussian SVM	172	37	45	126	0.74	0.82	0.78
15	Fine KNN	148	61	51	120	0.7	0.71	0.71
16	Medium KNN	161	48	43	128	0.75	0.77	0.76
17	Coarse KNN	121	88	20	151	0.88	0.58	0.72
18	Cosine KNN	169	40	59	112	0.65	0.81	0.74
19	Cubic KNN	161	48	44	127	0.74	0.77	0.76
20	Weighted KNN	152	57	32	139	0.81	0.73	0.77
21	Ensemble Boosted Trees	152	57	44	127	0.74	0.73	0.73
22	Ensemble Bagged Trees	170	39	46	125	0.73	0.81	0.78
23	Ensemble Subspace Discriminant	167	42	45	126	0.74	0.8	0.77
24	Ensemble Subspace KNN	169	40	64	107	0.63	0.81	0.73
25	Ensemble RUSBoosted Trees	143	66	46	125	0.73	0.68	0.71
26	EMPIRICAL SCORING ALGORITHM	47	162	5	166	0.97	0.23	0.56

Table 5 Modeling results for Scenario 2.

ID	Model Name	TN	FP	FN	TP	Sensitivity	Specificity	Accuracy
1	Fine Tree	149	60	58	113	0.66	0.71	0.69
2	Medium Tree	149	60	58	113	0.66	0.71	0.69
3	Coarse Tree	140	69	58	113	0.66	0.67	0.67
4	Linear Discriminant	153	56	64	107	0.63	0.73	0.68
5	Quadratic Discriminant	0	0	0	0	0	0	0
6	Logistic Regression	160	49	71	100	0.58	0.77	0.68

7	Gaussian Naïve Bayes	0	0	0	0	0	0	0
8	Kernel Naïve Bayes	163	46	73	98	0.57	0.78	0.69
9	Linear SVM	141	68	63	108	0.63	0.67	0.66
10	Quadratic SVM	150	59	58	113	0.66	0.72	0.69
11	Cubic SVM	150	59	58	113	0.66	0.72	0.69
12	Fine Gaussian SVM	150	59	59	112	0.65	0.72	0.69
13	Medium Gaussian SVM	155	54	59	112	0.65	0.74	0.70
14	Coarse Gaussian SVM	140	69	57	114	0.67	0.67	0.67
15	Fine KNN	87	122	26	145	0.85	0.42	0.61
16	Medium KNN	182	27	142	29	0.17	0.87	0.56
17	Coarse KNN	138	71	54	117	0.68	0.66	0.67
18	Cosine KNN	182	27	142	29	0.17	0.87	0.56
19	Cubic KNN	182	27	142	29	0.17	0.87	0.56
20	Weighted KNN	176	33	142	29	0.17	0.84	0.54
21	Ensemble Boosted Trees	148	61	59	112	0.65	0.71	0.68
22	Ensemble Bagged Trees	147	62	62	109	0.64	0.7	0.67
23	Ensemble Subspace Discriminant	158	51	65	106	0.62	0.76	0.69
24	Ensemble Subspace KNN	90	119	29	142	0.83	0.43	0.61
25	Ensemble RUSBoosted Trees	149	60	58	113	0.66	0.71	0.69
26	EMPIRICAL SCORING ALGORITHM	47	162	5	166	0.97	0.23	0.56

Table 6 Modeling results for Scenario 3.

ID	Model Name	TN	FP	FN	TP	Sensitivity	Specificity	Accuracy
1	Fine Tree	141	68	56	115	0.67	0.67	0.67
2	Medium Tree	141	68	49	122	0.71	0.67	0.69
3	Coarse Tree	138	71	60	111	0.65	0.66	0.66
4	Linear Discriminant	167	42	33	138	0.81	0.8	0.80
5	Quadratic Discriminant	166	43	69	102	0.6	0.79	0.71
6	Logistic Regression	175	34	31	140	0.82	0.84	0.83
7	Gaussian Naïve Bayes	180	29	24	147	0.86	0.86	0.86
8	Kernel Naïve Bayes	198	11	33	138	0.81	0.95	0.88

9	Linear SVM	185	24	34	137	0.8	0.89	0.85
10	Quadratic SVM	178	31	32	139	0.81	0.85	0.83
11	Cubic SVM	175	34	33	138	0.81	0.84	0.82
12	Fine Gaussian SVM	209	0	169	2	0.01	1.00	0.56
13	Medium Gaussian SVM	188	21	31	140	0.82	0.90	0.86
14	Coarse Gaussian SVM	185	24	30	141	0.82	0.89	0.86
15	Fine KNN	153	56	38	133	0.78	0.73	0.75
16	Medium KNN	159	50	27	144	0.84	0.76	0.80
17	Coarse KNN	150	59	14	157	0.92	0.72	0.81
18	Cosine KNN	177	32	41	130	0.76	0.85	0.81
19	Cubic KNN	159	50	26	145	0.85	0.76	0.80
20	Weighted KNN	154	55	25	146	0.85	0.74	0.79
21	Ensemble Boosted Trees	172	37	43	128	0.75	0.82	0.79
22	Ensemble Bagged Trees	166	43	45	126	0.74	0.79	0.77
23	Ensemble Subspace Discriminant	173	36	30	141	0.82	0.83	0.83
24	Ensemble Subspace KNN	177	32	32	139	0.81	0.85	0.83
25	Ensemble RUSBoosted Trees	154	55	50	121	0.71	0.74	0.72
26	EMPIRICAL SCORING ALGORITHM	47	162	5	166	0.97	0.23	0.56

Table 7 Modeling results For Scenario 4.

ID	Model Name	TN	FP	FN	TP	Sensitivity	Specificity	Accuracy
1	Fine Tree	138	71	53	118	0.69	0.66	0.67
2	Medium Tree	147	62	47	124	0.73	0.7	0.71
3	Coarse Tree	135	74	60	111	0.65	0.65	0.65
4	Linear Discriminant	147	62	53	118	0.69	0.7	0.7
5	Quadratic Discriminant	0	0	0	0	0	0	0
6	Logistic Regression	150	59	53	118	0.69	0.72	0.71
7	Gaussian Naïve Bayes	0	0	0	0	0	0	0
8	Kernel Naïve Bayes	153	56	61	110	0.64	0.73	0.69
9	Linear SVM	147	62	62	109	0.64	0.7	0.67
10	Quadratic SVM	144	65	54	117	0.68	0.69	0.69
11	Cubic SVM	139	70	56	115	0.67	0.67	0.67
12	Fine Gaussian SVM	153	56	72	99	0.58	0.73	0.66

13	Medium Gaussian SVM	152	57	64	107	0.63	0.73	0.68
14	Coarse Gaussian SVM	146	63	57	114	0.67	0.7	0.68
15	Fine KNN	146	63	90	81	0.47	0.7	0.6
16	Medium KNN	163	46	76	95	0.56	0.78	0.68
17	Coarse KNN	110	99	31	140	0.82	0.53	0.66
18	Cosine KNN	164	45	81	90	0.53	0.78	0.67
19	Cubic KNN	161	48	80	91	0.53	0.77	0.66
20	Weighted KNN	150	59	63	108	0.63	0.72	0.68
21	Ensemble Boosted Trees	144	65	45	126	0.74	0.69	0.71
22	Ensemble Bagged Trees	156	53	64	107	0.63	0.75	0.69
23	Ensemble Subspace Discriminant	151	58	59	112	0.65	0.72	0.69
24	Ensemble Subspace KNN	137	72	83	88	0.51	0.66	0.59
25	Ensemble RUSBoosted Trees	139	70	47	124	0.73	0.67	0.69
26	EMPIRICAL SCORING ALGORITHM	47	162	5	166	0.97	0.23	0.56

In the second experimental scenario, the Fine k-nearest neighbor (KNN) achieved the highest sensitivity of 85%. Other variants of KNN (i.e., Medium, Cosine, and Cubic KNNs) achieved classification accuracy values equal to that of the empirical scoring algorithm (i.e., 56%), highest specificity of 87% and a very low sensitivity of 17%. The Medium Gaussian SVM algorithm showed the highest classification accuracy of 70%. A noteworthy finding was that the Weighted KNN model showed the lowest classification accuracy of 54%, which was lower than that of the empirical scoring algorithm (56%). The results for Scenario 2 are further clarified in Table 5.

As shown in Table 6, the results for the third experimental scenario indicated that Coarse KNN achieved the highest sensitivity of 92%. Fine Gaussian SVM exhibited the highest specificity of 100% with the lowest sensitivity of 1% and classification accuracy equal to that of the empirical scoring algorithms (56%). Overall, for Scenario 3, Kernel Naïve Bayes appeared to be the best performing algorithm with the highest accuracy of 88%, specificity of 95%, and sensitivity of 81%.

Finally, in the fourth experimental scenario, Coarse KNN achieved the highest sensitivity of 82%, while Medium and Cosine KNNs showed the highest specificity of 78% each. Table 7 shows the experimental findings for this data modeling scenario.

Overall, variants of KNN and SVM were the best-performing models in all the scenarios, while the empirical model achieved better metrics, especially in Scenarios 2 and 4. The training sessions of Quadratic Discriminant and Gaussian Naïve Bayes algorithms failed in all the scenarios. Thus, zero values were recorded for these algorithms. Furthermore, the ML models achieved higher performances in Scenarios 1 and 3, both of which had the highest number of noncategorized input

parameters. Thus, the overall performance of the models was reduced in experimental scenarios 2 and 4, which are aligned to the clinical approach of parameter reduction.

4.3 Comparative Performance of the ML Models and the Empirical Scoring Algorithm Based on the Different Evaluation Metrics

The ML algorithms evaluated in the present study are not the most sophisticated ones used in other classification applications, but they have proved their merits in terms of predictive performance and efficiency. Thus, Figure 4(a)-(g) graphically presents the comparative performance of the multiple ML models and the empirical scoring algorithm based on the different evaluation metrics computed in Tables 4–7.

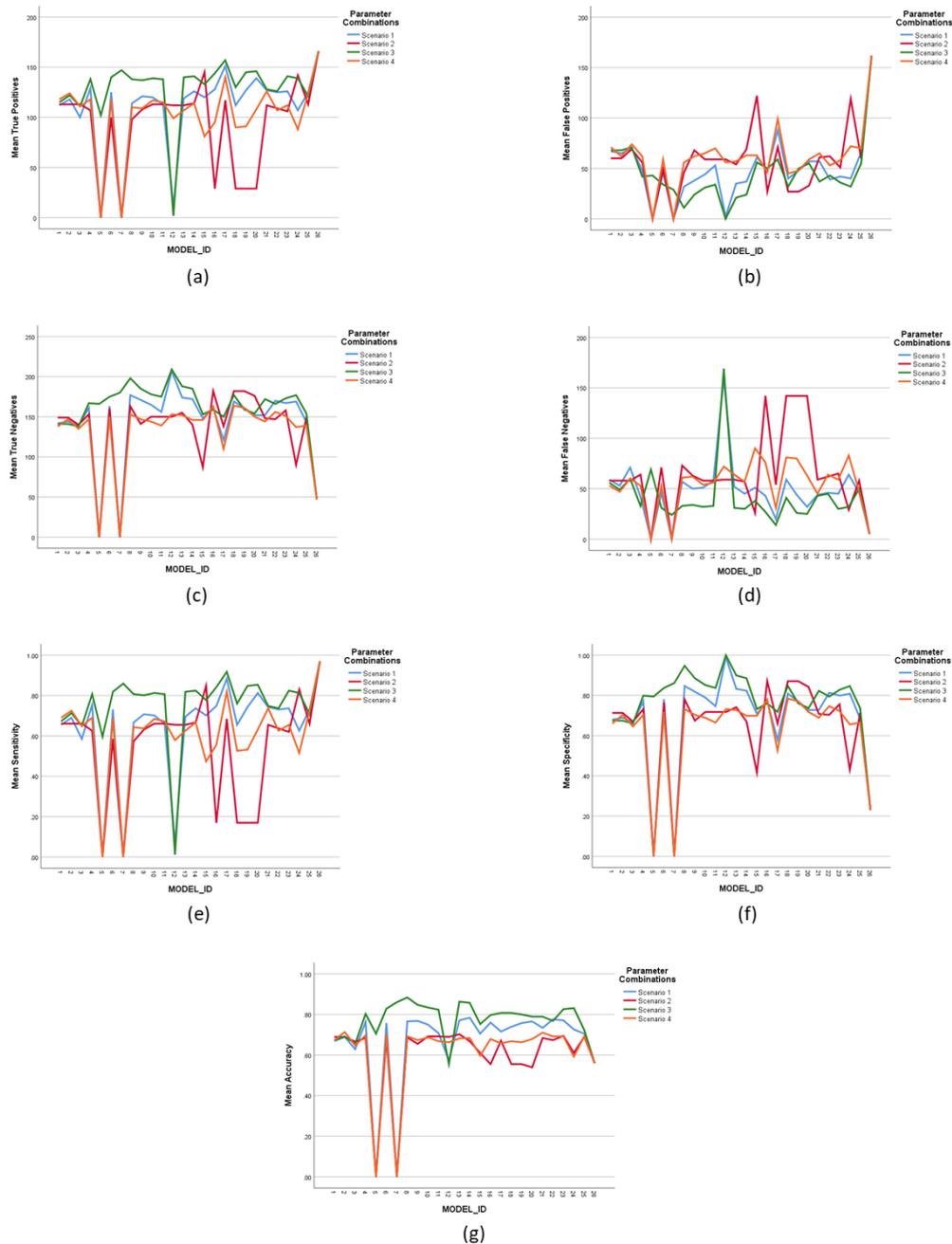


Figure 4 Comparative evaluation metrics based on the various experimental scenarios.

4.4 Discussion

The obtained results have provided several insights, especially regarding the effect of dimensionality reduction on model performance and result interpretation. First, the results obtained from the two experimental scenarios that used the untransformed input parameters (i.e., Scenarios 1 and 3) indicated better performance of the ML models in the predictive process based on the high evaluation metrics recorded. Closely related approaches and findings have shown that the inclusion of demographic parameters of severity level has improved the performance of the ML models [25, 30]. For instance, with the inclusion of severity level, Fine Gaussian SVM achieved an increased specificity of 1% between Scenario 1 and Scenario 3. This implies that apart from the main questionnaire items, demographic factors have a significant influence on the performance of the models. However, the customized rules in the empirical scoring algorithm do not consider demographic factors in the numerical quantification of ASD symptoms and in the final classification. Thus, the ML approach proved its merit in determining other influential factors that affect the predictive performance of the models. Another comparative analysis of classification accuracy verified that demographic factors influence the performance of models. Specifically, following the inclusion of the demographic factor of severity level, in Scenario 3, the Kernel Naïve Bayes classifier (with an accuracy of 88%, specificity of 95%, and sensitivity of 81%) achieved an increase of 10% in the classification accuracy over its performance in Scenario 1.

Comparative analysis between the original and transformed data provided insights into the effect of dimensionality reduction on the performance of the models and result interpretation. Even though the approach followed in the dimensionality reduction is based on the expert's knowledge, the results differ with respect to the evaluation metrics. Specifically, the performance of the models differed between the original and transformed data. For instance, the highest classification accuracy of 78% recorded for Scenario 1 declined by 8% and 7% after data transformations in Scenarios 2 and 4, respectively. This is in line with the assertion made by many studies on statistical irrelevancies that exist between the original and transformed data [44-47]. However, unlike data-centric approaches, the present approach preserved the clinical validity of the transformed data, despite a reduction in the performance of the models. This was noted in the comparative performance maintained by the empirical scoring algorithm across all the experimental scenarios. For instance, in Scenario 2, variants of KNN (i.e., Medium, Cosine, and Cubic KNN) achieved classification accuracy equal to that of the empirical scoring algorithm (56%), while the Weighted KNN model recorded the lowest classification accuracy of 54%, which was lower than that of the empirical scoring algorithm. Similarly, Fine Gaussian SVM despite achieving the highest specificity of 100%, its classification accuracy is equal to that of the empirical scoring algorithm (i.e., 56%). Other instances that could prove the worthiness of the transformation approach in preserving the clinical validity of the screening instrument are observed in the individual evaluation metrics highlighted in Figure 4(a)-(g), in which the empirical scoring algorithm achieved a very high sensitivity of 97% and very low FN rate (FN = 5). Thus, despite 160 FPs recorded from the empirical scoring algorithm, its clinical value is preserved because previous studies have indicated that the FP rate should be considered a critical factor while developing models for the medical diagnosis of ASD [48]. This is because the cost of misclassifying a non-autistic person (FP) is mild. Further diagnostic tests could correct such errors. Moreover, in medical diagnosis, FN rates bear a higher cost than FP rates. However, one of the critical implications of using the empirical scoring algorithm is that the high rates of FP and TP will

lead to implausible figures related to at-risk ASD cases. Overall, while the empirical scoring algorithms achieved outstanding performance in the correct classification of true ASD cases (sensitivity), the best-performing ML models outperformed the empirical method in the correct classification of non-ASD cases while achieving considerable classification accuracies.

5. Conclusion and Recommendations

5.1 Conclusion

Assessing the behavioral symptoms of patients with ASD by screening is a common preliminary stage for identifying people at risk of ASD and a crucial approach to fasten diagnostic referrals. Nonetheless, scoring autistic traits with the current screening instruments, such as the Autism-Spectrum Quotient (AQ) and ADI-R, relies solely on customized rules that have been associated with subjective interpretations. Thus, the trade-off in ASD screening and diagnostics studies is on improving the speed of the assessment processes and providing accurate and objective decisions. Previous studies have indicated that the merit of automated models based on ML techniques depends on accurate assessment systems from retrospective cases and controls. Recently, ML models for behavioral assessment of ASD have been broadly developed on the basis of a variety of pre-processed input data. Notably, previous studies mainly focused on rapid and accurate screening and diagnosis of ASD. However, to achieve rapidity of the process, various data selection and transformation techniques were used despite evidence of insufficiency of the reduced items and the inability of the transformation techniques to preserve the clinical validity of the screening and diagnostic instruments. Moreover, none of the previous studies investigated the sufficiency of the reduced parameters against the basis on which clinicians diagnose ASD. Consequently, clinical validity and real-life applicability of the ML models are at stake despite the superior evaluation metrics recorded in preceding studies. The performance of the models was evaluated based on the metrics of specificity, accuracy, sensitivity, and other variables. In essence, the multitudes of challenges for rapid and accurate ASD assessment are yet to be resolved by the preceding ML approaches. In the present study, ML was applied to behavioral screening and diagnosis of ASD by using a novel procedure that comprises a few behavioral features and preserves the clinical validity of the assessment instrument. Consequently, comparative analyses were performed between the empirical algorithm of the ASD screening instrument and multiple ML models. The study findings revealed the possibility of developing ML-based ASD assessment systems with excellent classification accuracy and adhering to the conceptual knowledge used in the clinical assessment of ASD. In the present study, an ML model based on Kernel Naïve Bayes was found to be the best-performing model, with a classification accuracy of 88%. The study findings and approaches pave the way for developing clinically valid ML-based systems that clinicians, parents, and other stakeholders can rely on for cost-effective screening and diagnosis of ASD symptoms.

5.2 Recommendations

Although several studies have demonstrated the use of the ML approach to assess ASD, future studies should establish the clinical relevance of the data-centric approaches and readjust the scientific use of the assessment instruments. Accordingly, future studies should investigate the best practices of scale development and feature reduction in line with the professional basis of ASD

diagnosis in categorizing and evaluating the clinical validity of robust ML models. Moreover, vital recommendations based on the findings of the present study were proposed based on the utilization of different experimental scenarios. Specifically, the best-performing ML models could be embedded in any ASD assessment app on the basis of the parameters used in the four data scenarios. In the first scenario, the ML-enabled ASD assessment app can have at most 30 input parameters. Although this scenario cannot streamline the parameters, the cost of implementation will be cheaper than that of the commonly used instruments such as SRS and ADOS, which have 65 and 93 items, respectively. Comparative analysis of the performance of the superior ML model against the empirical scoring algorithm indicated that among the key benefits of implementing the ML model is its excellent 72% increase in the TN rate over the 23% recorded for the empirical algorithm. Similarly, implementing the superior ML model of Scenario 3 could translate to the same benefits achievable in Scenario 1. However, implementing ML models with fewer input parameters can lead to reduced cost of the physical gadgets required and an improvement in the speed of administering the assessment tool. Specifically, implementing the superior models in Scenario 2 or Scenario 4 can provide an ML-embedded ASD screening app with at most eight input parameters with an overhead of implementing the empirical feature transformation rules. A comparative analysis between the ML models and the empirical scoring algorithm indicated that the best-performing ML model in Scenario 2 (i.e., Medium Gaussian SVM) achieved a 14% increase in classification accuracy over the empirical scoring algorithms. Moreover, despite having fewer items, the best-performing model in Scenario 4 outperformed the empirical scoring algorithm with increased accuracy and sensitivity of 15% and 47%, respectively. Another vital recommendation is concerning the present dimensions of the data collection instrument used in the present study. Future studies should look at the possibility of redesigning the data collection instrument and improving its scientific robustness on a behavioral scale. Recommendable approaches to categorize and establish valid dimensions from CDHSEBA include principal component analysis. Furthermore, future studies should implement enhanced instruments with more complex and robust algorithms, as well as some of the optimization techniques demonstrated in previous studies. Moreover, the visibility of the clinical validity of the proposed approach will enable clinicians to trust the worthiness of the evaluation metrics recorded. The present study is limited by the few cases sampled in the data collection stage and the possible factors that might have influenced the data responses. Other limitations include the use of a novel instrument that is not validated by multiple studies similar to other reputable scales employed in ASD assessment. Future studies should consider applying this research approach to the gold standard scales.

Author Contributions

Each of the authors has contributed equally to conceptualizing, drafting, and completing this research work.

Competing Interests

The authors have declared that no competing interests exist.

Additional Material

The following additional material is uploaded at the page of this paper.

1. Appendix A

References

1. Maenner MJ, Shaw KA, Bakian AV, Bilder DA, Durkin MS, Esler A, et al. Prevalence and characteristics of autism spectrum disorder among children aged 8 years—Autism and developmental disabilities monitoring network, 11 sites, United States, 2018. *MMWR Surveill Summ.* 2021; 70: 1-16.
2. Penner M, Anagnostou E, Andoni LY, Ungar WJ. Systematic review of clinical guidance documents for autism spectrum disorder diagnostic assessment in select regions. *Autism.* 2018; 22: 517-527.
3. Baio J, Wiggins L, Christensen DL, Maenner MJ, Daniels J, Warren Z, et al. Prevalence of autism spectrum disorder among children aged 8 years—Autism and developmental disabilities monitoring network, 11 sites, United States, 2014. *MMWR Surveill Summ.* 2018; 67: 1-23.
4. Case-Smith J, Weaver LL, Fristad MA. A systematic review of sensory processing interventions for children with autism spectrum disorders. *Autism.* 2015; 19: 133-148.
5. Chauhan A, Sahu JK, Jaiswal N, Kumar K, Agarwal A, Kaur J, et al. Prevalence of autism spectrum disorder in indian children: A systematic review and meta-analysis. *Neurol India.* 2019; 67: 100.
6. Durkin MS, Elsabbagh M, Barbaro J, Gladstone M, Happe F, Hoekstra RA, et al. Autism screening and diagnosis in low resource settings: Challenges and opportunities to enhance research and services worldwide. *Autism Res.* 2015; 8: 473-476.
7. Matson JL, Konst MJ. Early intervention for autism: Who provides treatment and in what settings. *Res Autism Spectr Disord.* 2014; 8: 1585-1590.
8. Campbell K, Carpenter KLH, Espinosa S, Hashemi J, Qiu Q, Tepper M, et al. Use of a digital modified checklist for autism in toddlers – Revised with follow-up to improve quality of screening for autism. *J Pediatr.* 2017; 183: 133-139.e1.
9. Shahamiri SR, Thabtah F. Autism AI: A new autism screening system based on artificial intelligence. *Cognit Comput.* 2020; 12: 766-777.
10. Thabtah F. An accessible and efficient autism screening method for behavioural data and predictive analyses. *Health Informatics J.* 2019; 25: 1739-1755.
11. Wingfield B, Miller S, Yogarajah P, Kerr D, Gardiner B, Seneviratne S, et al. A predictive model for paediatric autism screening. *Health Informatics J.* 2020; 26: 2538-2553.
12. Cavus N, Lawan AA, Ibrahim Z, Dahiru A, Tahir S, Abdulrazak UI, et al. A systematic literature review on the application of machine-learning models in behavioral assessment of autism spectrum disorder. *J Pers Med.* 2021; 11: 299.
13. Sekaran K, Sudha M. Predicting autism spectrum disorder from associative genetic markers of phenotypic groups using machine learning. *J Ambient Intell Humaniz Comput.* 2021; 12: 3257-3270.
14. Thabtah F. Machine learning in autistic spectrum disorder behavioral research: A review and ways forward. *Inform Health Soc Care.* 2019; 44: 278-297.

15. Frazier TW, Strauss M, Klingemier EW, Zetzer EE, Hardan AY, Eng C, et al. A meta-analysis of gaze differences to social and nonsocial information between individuals with and without autism. *J Am Acad Child Adolesc Psychiatry*. 2017; 56: 546-555.
16. Guillon Q, Hadjikhani N, Baduel S, Rogé B. Visual social attention in autism spectrum disorder: Insights from eye tracking studies. *Neurosci Biobehav Rev*. 2014; 42: 279-297.
17. Kowallik AE, Schweinberger SR. Sensor-based technology for social information processing in autism: A review. *Sensors (Basel)*. 2019; 19: 4787.
18. Negin F, Ozyer B, Agahian S, Kacdioglu S, Ozyer GT. Vision-assisted recognition of stereotype behaviors for early diagnosis of autism spectrum disorders. *Neurocomputing*. 2021; 446: 145-155.
19. Torres EB, Rai R, Mistry S, Gupta B. Hidden aspects of the research ADOS are bound to affect autism science. *Neural Comput*. 2020; 32: 515-561.
20. Bellesheim KR, Cole L, Coury DL, Yin L, Levy SE, Guinnee MA, et al. Family-driven goals to improve care for children with autism spectrum disorder. *Pediatrics*. 2018; 142: e20173225.
21. Duda M, Kosmicki JA, Wall DP. Testing the accuracy of an observation-based classifier for rapid detection of autism risk. *Transl Psychiatry*. 2015; 5: e556.
22. Kosmicki JA, Sochat V, Duda M, Wall DP. Searching for a minimal set of behaviors for autism detection through feature selection-based machine learning. *Transl Psychiatry*. 2015; 5: e514.
23. Küpper C, Stroth S, Wolff N, Hauck F, Kliewer N, Schad-Hansjosten T, et al. Identifying predictive features of autism spectrum disorders in a clinical sample of adolescents and adults using machine learning. *Sci Rep*. 2020; 10: 4805.
24. Wall DP, Dally R, Luyster R, Jung JY, Deluca TF. Use of artificial intelligence to shorten the behavioral diagnosis of autism. *PLoS One*. 2012; 7: e43855.
25. Usta MB, Karabekiroglu K, Sahin B, Aydin M, Bozkurt A, Karaosman T, et al. Use of machine learning methods in prediction of short-term outcome in autism spectrum disorders. *Psychiatr Clin Psychopharmacol*. 2019; 29: 320-325.
26. Wall DP, Kosmicki J, Deluca TF, Harstad E, Fusaro VA. Use of machine learning to shorten observation-based screening and diagnosis of autism. *Transl Psychiatry*. 2012; 2: e100.
27. Pratama TG, Hartanto R, Setiawan NA. Machine learning algorithm for improving performance on 3 AQ-screening classification. *Commun Sci Technol*. 2019; 4: 44-49.
28. Thabtah F, Kamalov F, Rajab K. A new computational intelligence approach to detect autistic features for autism screening. *Int J Med Inform*. 2018; 117: 112-124.
29. Thabtah F, Abdelhamid N, Peebles D. A machine learning autism classification based on logistic regression analysis. *Health Inf Sci Syst*. 2019; 7: 12.
30. Levy S, Duda M, Haber N, Wall DP. Sparsifying machine learning models identify stable subsets of predictive features for behavioral detection of autism. *Mol Autism*. 2017; 8: 65.
31. Baadel S, Thabtah F, Lu J. A clustering approach for autistic trait classification. *Inform Health Soc Care*. 2020; 45: 309-326.
32. Goel N, Grover B, Anuj, Gupta D, Khanna A, Sharma M. Modified grasshopper optimization algorithm for detection of autism spectrum disorder. *Phys Commun*. 2020; 41: 101115.
33. Bone D, Bishop SL, Black MP, Goodwin MS, Lord C, Narayanan SS. Use of machine learning to improve autism screening and diagnostic instruments: Effectiveness, efficiency, and multi-instrument fusion. *J Child Psychol Psychiatry*. 2016; 57: 927-937.

34. Suresh Kumar R, Renugadevi M. Differential evolution tuned support vector machine for autistic spectrum disorder diagnosis. *Int J Recent Technol Eng.* 2019; 8: 3861-3870.
35. Akter T, Satu MS, Khan MI, Ali MH, Uddin S, Lio P, et al. Machine learning-based models for early stage detection of autism spectrum disorders. *IEEE Access.* 2019; 7: 166509-166527.
36. Thabtah F, Peebles D. A new machine learning model based on induction of rules for autism detection. *Health Informatics J.* 2020; 26: 264-286.
37. Puerto E, Aguilar J, López C, Chávez D. Using multilayer fuzzy cognitive maps to diagnose autism spectrum disorder. *Appl Soft Comput.* 2019; 75: 58-71.
38. Duda M, Ma R, Haber N, Wall D. Use of machine learning for behavioral distinction of autism and ADHD. *Transl psychiatry.* 2016; 6: E732.
39. Duda M, Haber N, Daniels J, Wall DP. Crowdsourced validation of a machine-learning classification system for autism and ADHD. *Transl Psychiatry.* 2017; 7: e1133.
40. Song DY, Kim SY, Bong G, Kim JM, Yoo HJ. The use of artificial intelligence in screening and diagnosis of autism spectrum disorder: A literature review. *Soa Chongsonyon Chongsin Uihak.* 2019; 30: 145-152.
41. Trevethan R. Sensitivity, specificity, and predictive values: Foundations, pliabilitys, and pitfalls in research and practice. *Front Public Health.* 2017; 5: 307.
42. Appakaya SB, Sankar R, Ra IH. Classifier comparison for two distinct applications using same data. *Proceedings of the 9th International Conference on Smart Media and Applications; 2020 September 17-19; Jeju, Republic of Korea.* New York: Association for Computing Machinery.
43. Lalkhen AG, McCluskey A. Clinical tests: Sensitivity and specificity. *Contin Educ Anaesth Crit Care Pain.* 2008; 8: 221-223.
44. Curtis AE, Smith TA, Ziganshin BA, Elefteriades JA. The mystery of the Z-score. *Aorta (Stamford).* 2016; 4: 124-130.
45. Feng C, Wang H, Lu N, Chen T, He H, Lu Y, et al. Log-transformation and its implications for data analysis. *Shanghai Arch Psychiatry.* 2014; 26: 105-109.
46. Lapteacru I. On the consistency of the Z-score to measure the bank risk. *Bordeaux: University of Bordeaux; 2016.* Available from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2787567.
47. Wiesen JP. Benefits, drawbacks, and pitfalls of Z-score weighting. *Proceedings of 30th Annual IPMAAC Conference; 2006 June 27; Las Vegas, NV, USA.* Available from: <https://appliedpersonnelresearch.com/papers/Wiesen.2006.IPMAAC.z-score.weighting.pdf>.
48. Alahmari F. A comparison of resampling techniques for medical data using machine learning. *J Inf Knowl Manag.* 2020; 19: 2040016.



Enjoy *OBM Neurobiology* by:

1. [Submitting a manuscript](#)
2. [Joining in volunteer reviewer bank](#)
3. [Joining Editorial Board](#)
4. [Guest editing a special issue](#)

For more details, please visit:

<http://www.lidsen.com/journals/neurobiology>